Water Practice & Technology

IWA PUBLISHING

Water Practice & Technology Vol 17 No 12, 2499 doi: 10.2166/wpt.2022.146

Investigation of scarce input data augmentation for modelling nitrogenous compounds in South African rivers

Christopher Dumisani Mahlathi[a,*], Josefine Wilms[b] and Isobel Brink[c]

[a] Council for Scientific and Industrial Research, P.O. Box 320, Stellenbosch 7599, South Africa
[b] Deutsches GeoForschungs Zentrum, Claude-Dornierstr. 1, Gebäude 401, Raum 1.05, Weßling 82234, Germany
[c] Department of Civil Engineering, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa
*Corresponding author. E-mail: cdmahlathi@csir.co.za; cdmahlathi@gmail.com

ABSTRACT

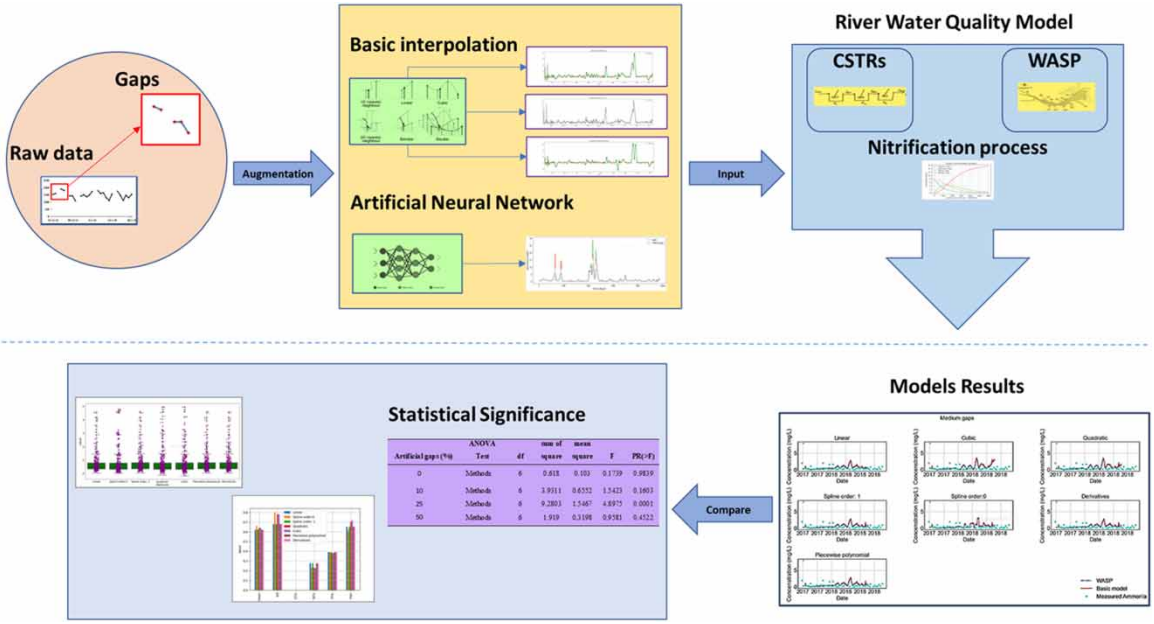
In this study, basic interpolation and machine learning data augmentation were applied to scarce data used in Water Quality Analysis Simulation Programme (WASP) and Continuous Stirred Tank Reactor (CSTR) that were applied to nitrogenous compound degradation modelling in a river reach. Model outputs were assessed for statistically significant differences. Furthermore, artificial data gaps were introduced into the input data to study the limitations of each augmentation method. The Python Data Analysis Library (Pandas) was used to perform the deterministic interpolation. In addition, the effect of missing data at local maxima was investigated. The results showed little statistical difference between deterministic interpolation methods for data augmentation but larger differences when the input data were infilled specifically at locations where extrema occurred.

Key words: data enhancement, interpolation, machine learning, nitrogenous compounds, water quality modelling

HIGHLIGHTS

- Basic interpolation methods did not produce statistically significant differences in augmented datasets.
- Increasing the gaps yielded greater differences between augmented datasets.
- ML methods on real and artificial gaps produced acceptable results.
- No significant differences between the WASP and Basic Model on real and artificial input.
- Difference between the WASP and Basic Model on real and artificial input.

GRAPHICAL ABSTRACT



INTRODUCTION

Fresh water is a scarce resource in South Africa, with an average annual daily rainfall of 490 mm (WWF-SA 2016) being around half the global average. A recent local annual climate report shows that precipitation patterns have remained unchanged in South Africa (SAWS 2020). There has therefore not been any increase in rainfall while the demand for water increased steadily. In addition to water scarcity, South African freshwater resources are subjected to strain through pollution from underperforming water treatment facility effluents. Eutrophication, a result of excess phosphoric and nitrogenous nutrients in a river system, has been previously highlighted by van Ginkel (2011) and Harding (2015) as a serious problem in the country. Therefore, effective water resource management is crucial to managing the quantity and quality of available in-demand levels.

Water resource management solutions require a reliable representation of the state of water in the regions of the country where water is monitored and managed. This can be achieved through the use of established sampling networks that record the quantity and quality of available water for various uses. Numerical models can be further used to complement actual measurements to achieve a similar objective.

Numerical models can describe processes (such as geo-hydrology, climate properties, sources, sinks) recorded in the field to provide a more reliable and accurate representation. This is possible when adequate measured boundary condition data as well as data for calibration and validation are available. When data are scarce, augmentation techniques can be applied to adjust the recorded information for the model input, which is necessary for a more accurate simulation of the environment. Data availability can therefore limit the choice of reasonable modelling approaches (including items of structure, complexity, and spatial and temporal resolution) that can be applied (Slaughter *et al.* 2017).

Spatial and temporal resolution are examples of necessary requirements for the model study that seeks to simulate a dynamic process such as nitrification in rivers. This study focuses on the impact of augmented model input data generated by applying simple (basic interpolation) and advanced (artificial neural networks (ANNs)) augmentation methods applicable to modelling nitrogenous compounds in the river system.

In instances where the input data are of unsatisfactory resolution; data augmentation techniques may be applied for gap filling and disaggregation to meet the model requirements (Baffaut *et al.* 2015). Examples of recent successful applications of data augmentation through interpolation to water quality modelling can be obtained in the literature (see Yang *et al.* 2020 and Kim *et al.* 2021). Blöschl & Sivapalan (1995) details how upscaling and downscaling of data can be used to align the scales of available data with model requirements. In this context, upscaling refers to the transfer of information from a small scale to a large scale and downscaling refers to transferring information to a small scale. Downscaling consists of two steps (disaggregation and singling out) to transfer information to a smaller scale (Blöschl & Sivapalan 1995). Input hydrodynamic data may be disaggregated from monthly to daily flows to accommodate the variation in daily concentration during water quality modelling. This was demonstrated when flow duration curves and mathematical relationships were applied to flow data for the WQSAM model to study water quality (see Slaughter (2017)). Disaggregation in hydrology is one of the steps used for downscaling hydrology data to meet the scales required for meeting the modelling objective.

Time series for hydrological processes can also be generated using either deterministic (basic interpolation methods) or stochastic (ANNs) models. Koutsoyiannis *et al.* (2008) detail the differences in the application of stochastic models against deterministic models. The important finding in this study is that good stochastic models are those that are linked with an understanding of the natural behaviours of the system. Furthermore, deterministic models such as the analogue model can be a good simplistic analytical tool, however, good prediction from this model does not necessarily represent consistency with natural processes. Since this study focuses on nitrogenous compounds which can be closely linked to hydrogen cycle processes through river stream flow, some stochastic models may not be ideal for generating time-series analysis depending on whether autocorrelation structure is for short range dependence or long-range dependence (Dimitriadis *et al.* 2021).

There is a variety of gap-filling methods in the literature that have been applied successfully in water quality modelling studies. Table 1 lists studies on this topic.

In this study, different augmentation methods applied towards modelling a nitrogenous compound in the river system were investigated. This was done to inform augmentation method choice when dealing with scarce data. First, the impact of interpolation method choice as well as whether the level of gaps in the input data impacts the outcome of each augmentation method differently was investigated. Second, an advanced interpolation method

Table 1 | Gap-filling methods applied to water quality models

Method	Application	Source
Singular Spectrum Analysis (SSA)	Gap-filling hydrological data	Sandoval *et al.* (2016)
Using Principal Components Analysis and Inverse Distance Weighted (IDW) Interpolation	Spatial and temporal changes in surface water quality	Yang *et al.* (2020)
Delaunay and k-Nearest Neighbours (kNN)	Spatio-temporal analysis of river water quality parameters	Vizcaino *et al.* (2016)
Linear interpolation and regression methods	Estimation of decadal stream flow	Lee *et al.* (2016)
Statistical models and ANNs	Gap-filling techniques for river stage data	Luna *et al.* (2020)
ANNs	Augmentation of limited input data for water quality model or a lake	Kim *et al.* (2021)

(machine learning) was applied towards gap filling while exploring the best method for partition training and validation given limited data. Finally, the output of the models was compared for scenarios where different levels of artificial gaps are introduced in the input data.

The outcome of this study provides guidance towards augmenting data for water quality modelling under scarce data conditions, which are similar to real-world data availability in the study area. Improvement of model fit to the measured data is beyond the scope of this paper, however there is evidence that the selection of the appropriate machine learning method can translate to better model fit (see Rozos *et al.* 2022). This study is limited to observing the significance of the differences between the augmentation method output.

METHODS

Two scenarios were investigated. First, the effect of using deterministic interpolation for data augmentation. Second, the effect of using deterministic and machine learning interpolation to infill weekly gaps at locations where maxima occur in the ammonia input was also investigated.

Study area

The study area consists of a 5.9 km river reach of the Natalspruit river system in the Upper Vaal catchment in South Africa. The river reach receives effluent discharge from one wastewater treatment plant (WWTP). The river data sets include observed time-series concentration data at two sampling points located upstream and downstream of the WWTP discharge location. The observed data were collected at a weekly frequency between 2012 and 2020; with some gaps observed in certain years. Figure 1 shows the upstream and downstream sampling locations and the WWTP effluent discharge location.

Augmentation methods

Two data operations (gap filling and temporal disaggregation) were required to eliminate the gaps in the original dataset and to upscale the temporal resolution of the input data from weekly to daily frequency. This is to meet the model requirement for dynamic processes according to the recommendations by Moriasi *et al.* (2012).

Basic interpolation

The chosen basic interpolation methods (linear, quadratic, cubic, spline (first and second order), polynomial, piecewise polynomial, and derivatives) in the Python Pandas library (The pandas development team 2021) were applied to upscale data to a daily resolution. A detailed discussion of these interpolation methods can be found in Virtanen *et al.* (2020). The scope of this article only covers the application of each method; whereas method algorithm details can be found in the references listed in Table 2.

Each of these interpolation methods was applied to generate an augmented dataset with a daily frequency; as recommended by Baffaut *et al.* (2015) for modelling the dynamic processes that affect nitrogenous compounds in river systems. Upsampling in this study refers to when the frequency of the samples is increased such as from weekly to daily. For this to be done, the generation of a time series is required. Resampling, in this case, is required to transform the available irregular frequency of data to a regular frequency and to increase the

Figure 1 | Map showing the study area considered (Google Earth Pro 7.3.4, Natal Spruit, 26°15′55″S, 25°11′30″E, Maxar Technologies, February 2022).

Table 2 | Interpolation methods for augmenting missing data gaps

Method	Description	Source
Linear interpolation	Curve fitting method using linear polynomials to generate estimated data point within the range of a discrete set of known data points.	Siauw & Bayen (2015)
Quadratic	Interpolation using second-order polynomial to make interpolation for a function	Vandebogert (2017)
Derivatives	Interpolation using derivative information that is a hybrid of extrapolation to arbitrary order and linear interpolation.	Tugores & Tugores (2017)
Polynomial	Interpolation of a given data set by the polynomial of the lowest possible degree that passes through the points of the dataset	Zou *et al.* (2020)
Piecewise Cubic Hermite Interpolating Polynomial (PCHIP)	Spline interpolator where each piece is a third-degree polynomial specified in Hermite form	Barker & McDougall (2020)
Cubic spline	Interpolation is where the interpolant is a special type of piecewise polynomial called a spline.	László (2005)
S-linear	Spline interpolation of order 1	Virtanen *et al.* (2020)
Zero	Spline interpolation of order 0	Virtanen *et al.* (2020)

number of samples to create more data on which the neural network can be trained. It is important to note that resampling methods have the limitation of destroying the long-range dependence that appeared in the data.

Simulation models

Two different simulation models were included in this study.

Continuously Stirred Tank Reactor in series model

A basic Continuously Stirred Tank Reactor (CSTR) model was used to represent the simplest form of a completely mixed natural water body. Here, several CSTRs were placed in series to simply simulate river sections. This

method, detailed in Chapra (1997), solves the mass balance equation, which takes the form in Equation (1) below. For simplicity, only a feed-forward system was considered:

$$V\frac{dc}{dt} = W(t) - kV_c - vA_s c \tag{1}$$

In Equation (1), V represents the reactor volume, c is parameter concentration in the reactor, $W(t)$ represents the lumped loading, t is the time, k is the reaction rate constant, A_s is the cross-sectional area, and v is the flow velocity.

The reaction term on the mass balance equation represents the nitrification process in a river system. The process occurs in two reaction steps. Step 1 shown by Equation (2), *Nitrosomonas* bacteria convert ammonium ions to nitrite (Chapra 1997):

$$NH_4^+ + 1.5O_2 \rightarrow 2H^+ + H_2O + NO_2^- \tag{2}$$

Step 2, represented by Equation (3) *Nitrobacter* bacteria convert nitrite to nitrate

$$NO_2^- + 0.5O_2 \rightarrow NO_3^- \tag{3}$$

The oxygen required in the two steps can be determined as (Chapra 1997)

$$r_{on} = r_{oa} + r_{oi} = 4.57\ gO\ gN^{-1} \tag{4}$$

where r_{on} is the amount of oxygen consumed per unit mass of nitrogen in the total nitrification reaction. r_{oa} and r_{oi} are the total oxygen consumed due to nitrification of ammonia and nitrite, respectively. Usually first-order kinetics are assumed for modelling the nitrification process and the following Equations (5)–(8) as described in Chapra (1997) were included:

$$\frac{dN_0}{dt} = -k_{0a}N_0 \tag{5}$$

$$\frac{dN_a}{dt} = k_{0a}N_0 - k_{ai}N_a \tag{6}$$

$$\frac{dN_i}{dt} = k_{ai}N_a - k_{in}N_i \tag{7}$$

$$\frac{dN_n}{dt} = k_{in}N_i \tag{8}$$

where N is the parameter concentration and the subscripts o, a, i, and n denote organic, ammonium, nitrite, and nitrate, respectively. The oxygen deficit (D) balance can be computed as written in Equation (9):

$$\frac{dD}{dt} = r_{0a}k_{ai}N_a + r_{0i}k_{in}N_i - k_aD \tag{9}$$

These equations were solved using Python's fourth-order Runge-Kutta solver for numerical integration because it was simple to apply to the system of differential equations in this study. Ammonia concentration was computed on the selected checkpoints in the river reach.

Water Quality Analysis Simulation Programme

Water Quality Analysis Simulation Programme (WASP) software was additionally used to model the river reach. WASP (Wool *et al.* 2020) is an open-source dynamic compartment-modelling programme for aquatic systems, including both the water column and the underlying benthos. It was developed and distributed by the Environment Protection Agency in the United States (Wool *et al.* 2020). It allows the user to investigate 1-, 2-, and 3-dimensional systems as well as a variety of pollutant types and processes such as eutrophication. This programme was selected because it is a recognised software that is capable of modelling nitrogenous compounds in a river

system. Details about the development of this programme and capabilities can be obtained from the work of Wool *et al.* (2020).

Input data

Measured on-site data sets were available, but with varying spatial consistency, which required cleaning where data values were absent in addition to gap filling for missing data. Each data sample (upstream and downstream from the WWTP) showed varying information. Table 3 provides a list of raw data features.

Table 3 | Summary of measured ammonia data entries

		Ammonia concentration distribution (mg/L)				
Location	**Total entries**	**0–1**	**1–5**	**5–10**	**>10**	**Not a number**
Upstream samples	316	305	45	4	6	0
Downstream samples	343	173	160	24	4	0

Hydrodynamic data measurements on the reach were not available. Estimates of flow rates were derived using measurements from a nearby station at a reach with similar features (catchment size and climate) as recommended by Daggupati *et al.* (2015) when dealing with data scarcity. The available hydrodynamic data covered a full year (between 1 October 2017 and 1 October 2018). The reader is reminded here that the focus of this research was not to test model fit, but rather to investigate differences in model outcomes when using different water quality data augmentation methods. This data application was therefore accepted as a realistic representation of hydrodynamic data for the system.

Investigation design

As mentioned previously, the study area consisted of a single river reach with one WWTP discharging into the stream. The upstream boundary of the study consisted of observed water quality data with weekly temporal resolution for the period 2012–2020. The water quality parameters relevant to this study were monitored nitrogenous compounds (ammonia and nitrates). The data sets upstream of the WWTP were used as a boundary condition for the river models; the downstream observed data sets were used for output comparisons. The augmentation study focused on the impact that the boundary condition data had on the model outputs as observed at the downstream boundary. The focus was to discern whether applying different augmentation techniques would yield significant differences in the model outputs.

Basic interpolation method study

The input ammonia concentration data were divided into four categories:

- No gaps – the raw measured upstream boundary data as measured.
- Low gaps – raw data with 10% random artificial gaps.
- Medium gaps – raw data with 25% random artificial gaps.
- High gaps- raw data with 50% random artificial gaps.

Each data set was subjected to interpolation for the data gaps and upsampled to daily concentration data through applying the linear, quadratic, derivatives, cubic, piecewise polynomial, 1st order spline and 0th order spline interpolation methods. The augmented data were used as input to the Basic Model (CSTRs in series) and the WASP model in turn to simulate a 5.9 km long single reach river system for nitrogenous compounds with ammonia selected as a proxy parameter to nitrogenous compounds to represent the changes brought about through the nitrification process. The simulation models were driven by flowrates as estimated using transference as explained above. The output of each model for each input data set as generated through the use of the interpolation methods was compared to determine statistically significant differences.

Advance interpolation method (ANNs) study

As previously stated, two scenarios are investigated. First, we investigate the effect of using deterministic interpolation for data augmentation. Second, we investigate the effect of using deterministic and machine learning interpolation to infill weekly gaps at locations where maxima occur in the ammonia input.

The raw data samples were irregular (sampled at 5–7 days) and the data set size was less than the recommended fifty times the number of weights in the ANN (see Alwosheel *et al.* (2018)). To convert this to a data set on which a neural network could be trained, the data were upsampled to a daily frequency with a linear interpolation method. The distribution of the upsampled data before and after the log transform is shown in Figures 2–4.

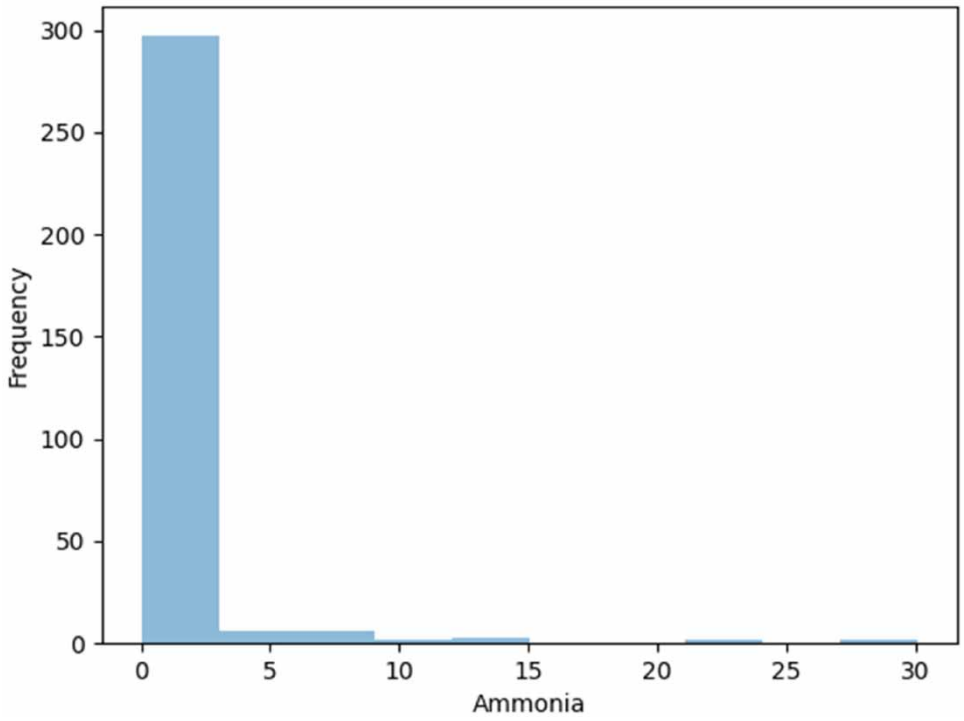


Figure 2 | Input ammonia data distribution.

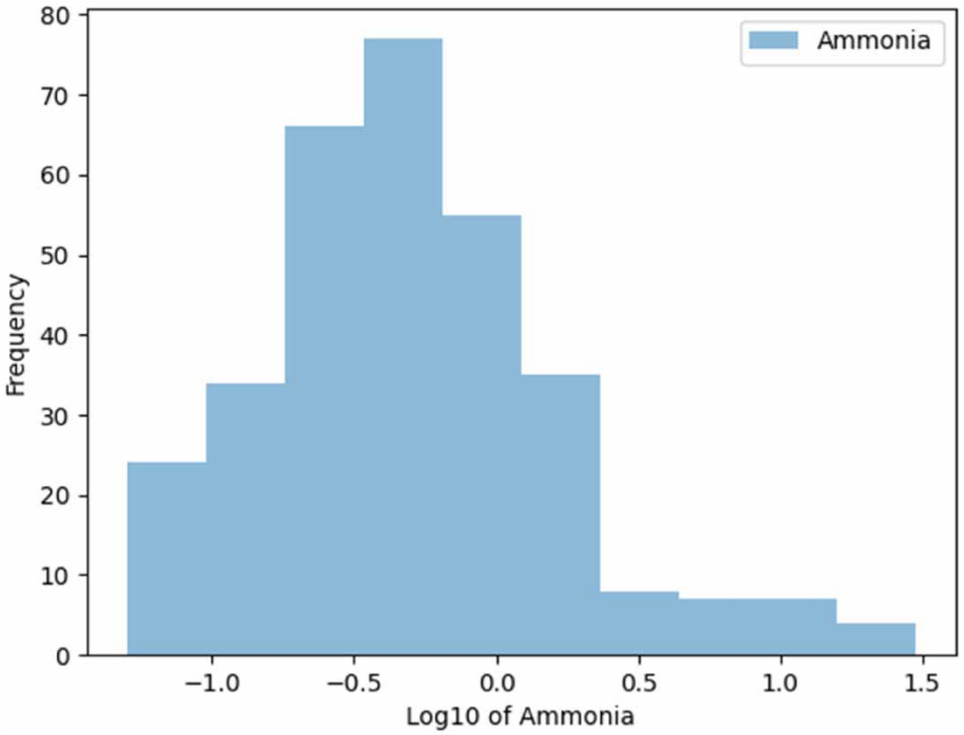


Figure 3 | Log-transformed input ammonia data.

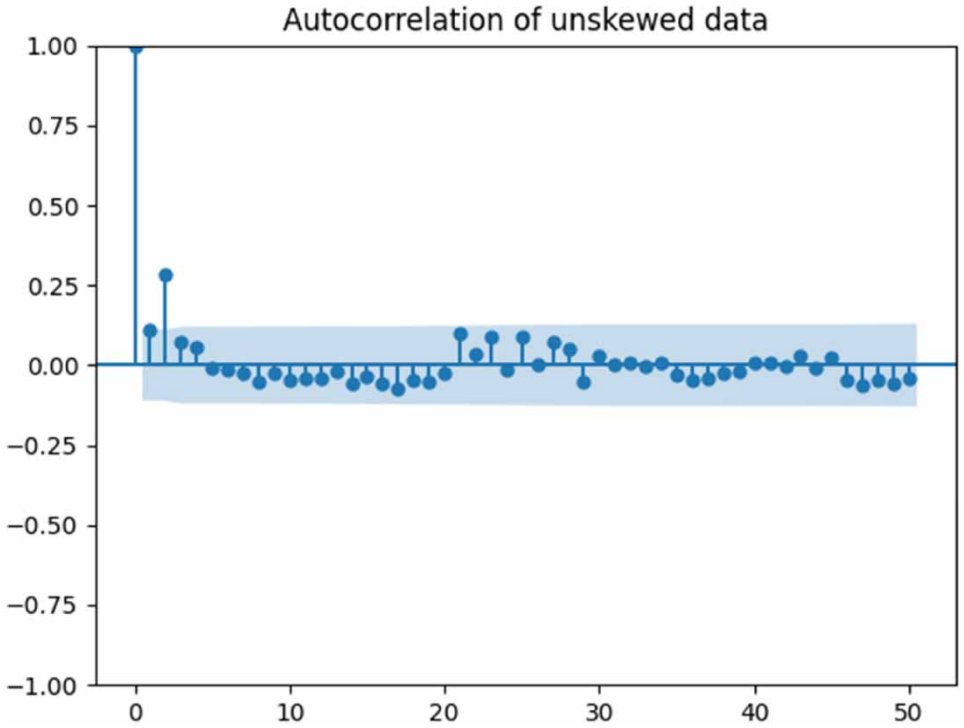


Figure 4 | Autocorrelation plot showing a linear dependence of ammonia data for 3 consecutive days and reduced correlations for lags larger than 3 days.

The log transform is needed here to prevent the minority of very large ammonia values to influence our model training too significantly: by applying a log transform action, the small and large ammonia values become comparable. Normalising the data generally speeds up learning and leads to faster convergence.

To analyse the dependence (or correlation) of consecutive measurements on each other, an autocorrelation plot was made of the upsampled data. Autocorrelation plots of the data are shown in Figure 4.

Figure 4 shows a significant correlation for a lag of 3 days or less. It should be noted that this can only be used as a guideline when selecting the number of timesteps to use to predict the next sequence of timesteps; since it only provides information on the linear- and not the non-linear relationship of the measurements. The pre-processing stage was then concluded by dividing the data into a training- and a testing set for which 70% of the upsampled data were used for training the neural network and 30% was used for testing.

Model architecture and optimisation

For background information on neural networks the reader is referred to Mehlig (2021). The neural network used here consisted of a single hidden layer. It was optimised to infill a week of missing ammonia values by using 7 days (1 week) measurements before and after the gap as input to the ANN model. The optimal numbers of input nodes and training epochs were determined by training on the training dataset for 10–500 input nodes and, for each of these, node selections of 10–100 epochs were used. For each of these scenarios, 30% of the data were used for validation and model performance was evaluated by calculating the mean squared error (MSE) on the validation data set. The number of epochs and nodes that resulted in the model with the smallest MSE on the validation data set was then trained on the entire training data set and saved. To show that the model was not over-trained, the MSE curves for training and validation are shown in Figure 5.

The validation plot did not increase towards the training line, which is an indication that the model was not over-trained.

Model testing

The optimal saved model was then applied to the test set and the prediction vs. target value, for each of the seven timesteps within the gap, are shown in Figure 6. These artificial gaps were created specifically at locations where local maxima occur within the upsampled data.

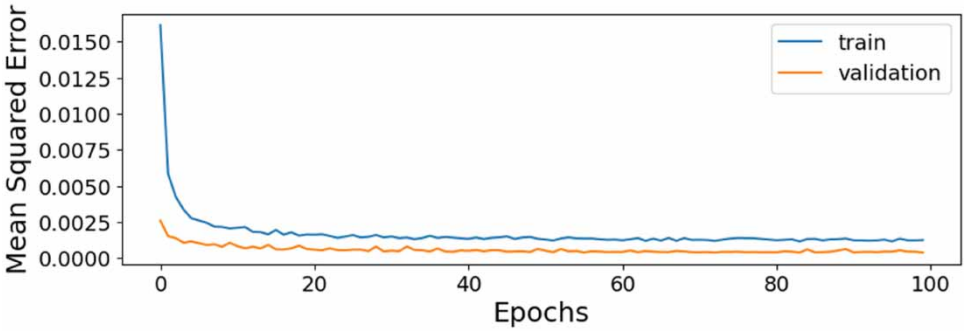


Figure 5 | MSE plot showing the relationship between training and validation sets.

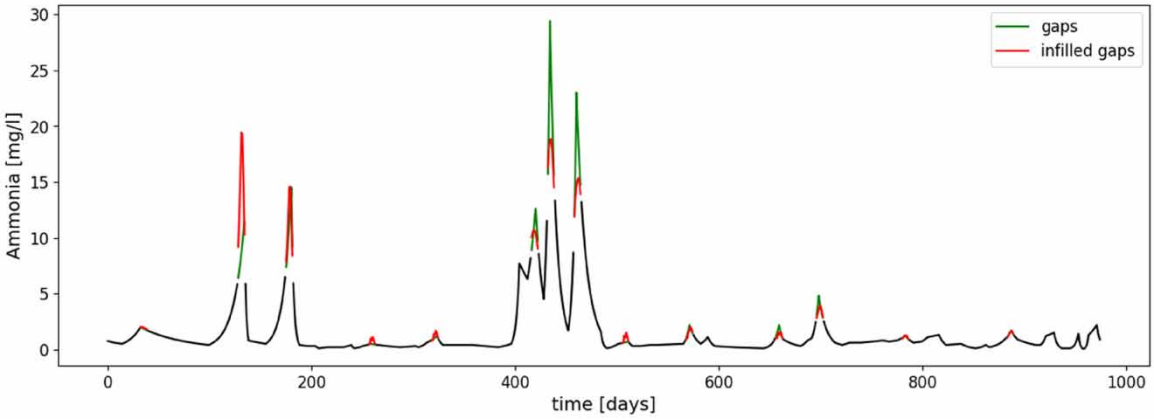


Figure 6 | Comparison of the actual data (in green) and the predictions, made by the ML model (in red). Please refer to the online version of this paper to see this figure in colour: http://dx.doi.org/10.2166/wpt.2022.146.

To visualise the effectiveness of the model to predict gaps that contain local maxima, artificial 7-day gaps were created within the test set by selecting positions that contain local maxima.

These gaps were then also infilled with the deterministic statistical interpolation methods and the interpolated results were used as input to the water quality model. The outputs of the water quality model for each of these input scenarios are shown in Figure 7.

In this case, the machine learning method seems to predict high peaks compared to the rest of the methods (linear, Piecewise Cubic Hermite Interpolating Polynomial (PCHIP)) except for the polynomial method at the local maxima.

Result evaluation method

This part of the investigation was done to determine the effect of time-series data disaggregation by interpolation on the boundary of a river simulation method. The design layout involved simulating the selected model for each of the interpolation methods. The output of the models for each method was compared, within a model, and then between the models.

To evaluate the comparisons between the models, statistical techniques were applied to determine if the difference between the simulated output of the interpolation model was statistically significant. If the difference is significant, this will indicate that the choice of an interpolation method has an impact on the outcome of a simulation model. Therefore, the choice of the augmentation model should be considered a crucial variable when a model is developed for a scarce data system.

The opposite of this outcome would suggest that for this dataset, the choice of an interpolation method has no significant impact on the output of the models. This would imply that when setting up a model with inadequate data, the choice of interpolation method from the list discussed in this study, does not make a significant difference to the model output.

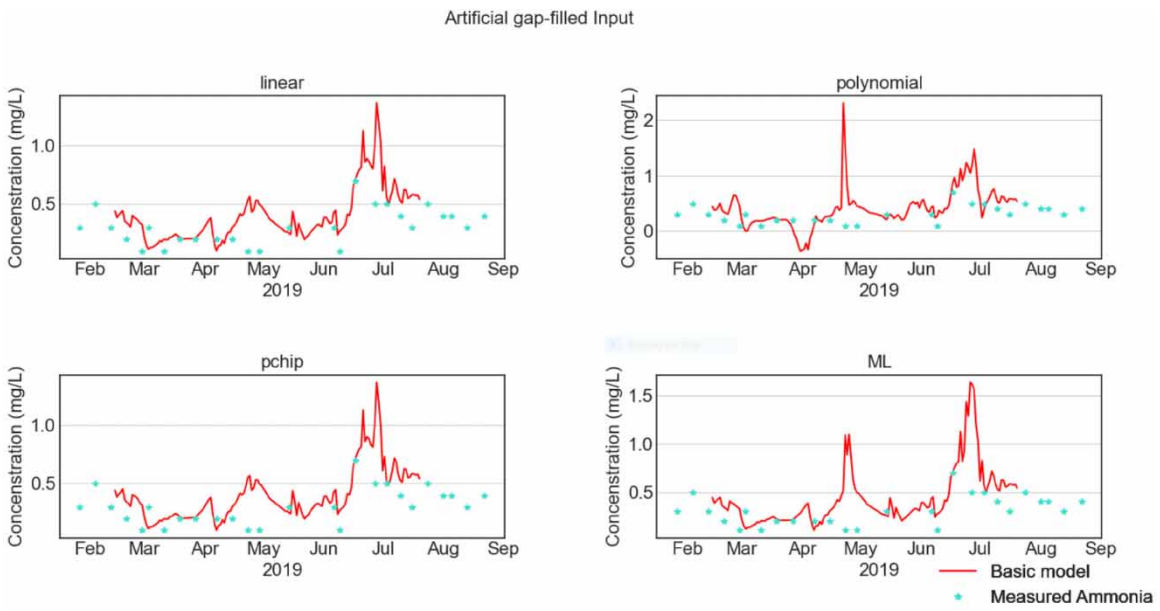


Figure 7 | Basic Model output plots for input data filled with basic interpolations (linear, polynomial, and PCHIP) and ML.

To quantify this, two statistical methods were applied to determine the differences caused by the model boundary interpolation methods. The differences between the interpolated boundary conditions were determined using quantitative statistics and inferential statistics. For this study, the *T*-test and the ANOVA test were applied to the model results.

A *T*-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups (model output from different interpolation methods), while the ANOVA test does the same for more than two groups. The null hypothesis for both tests assumes that there is a significant difference between the groups, which is an assumption made about the groups. The *T*-test and ANOVA test used t values and F values, respectively, to determine whether the null hypotheses pass or fail. A probability value (*p*-value) is also calculated for each test that suggests whether the hypothesis should be accepted (chosen $p < 0.05$).

The results of the model simulation for the Basic Model and the WASP model were analysed using the *T*-test and the ANOVA test. This was done to evaluate how each method compares for both models. This was done to determine the differences between all the methods in each model. This was followed by the *T*-test for methods output comparison across models (Basic Model output against WASP model output) to evaluate the differences between the models.

RESULTS AND DISCUSSION

Descriptive comparison of model outputs

Basic Model

Figure 8 shows descriptive statistics of the downstream model outputs per basic interpolation method. Figure 8 indicates differences in model output per interpolation method. The significance of which is further investigated below. Here, the means of each method showed a difference of less than 0.1 mg/L, a rather small amount indicating that the differences in outputs may have been insignificant across interpolation method applications. Furthermore, the standard deviation in output values (denoted by std) for spline 0 order, quadratic, and cubic resulted in similar, but larger when compared to the other method outputs, standard deviations in the range of 0.80–0.85; whereas the other methods resulted in similar values of approximately 0.7. The 25% interquartile ranges (IQRs) of the model outputs resulted in larger values for the quadratic and cubic methods when compared to the other interpolation methods investigated. However, the 50 and 75% IQRs once more resulted in similar output values across interpolation methods. It was also noted that the 25% IQR produced negative values for the quadratic and cubic methods. The reasons for which are unclear at this stage.

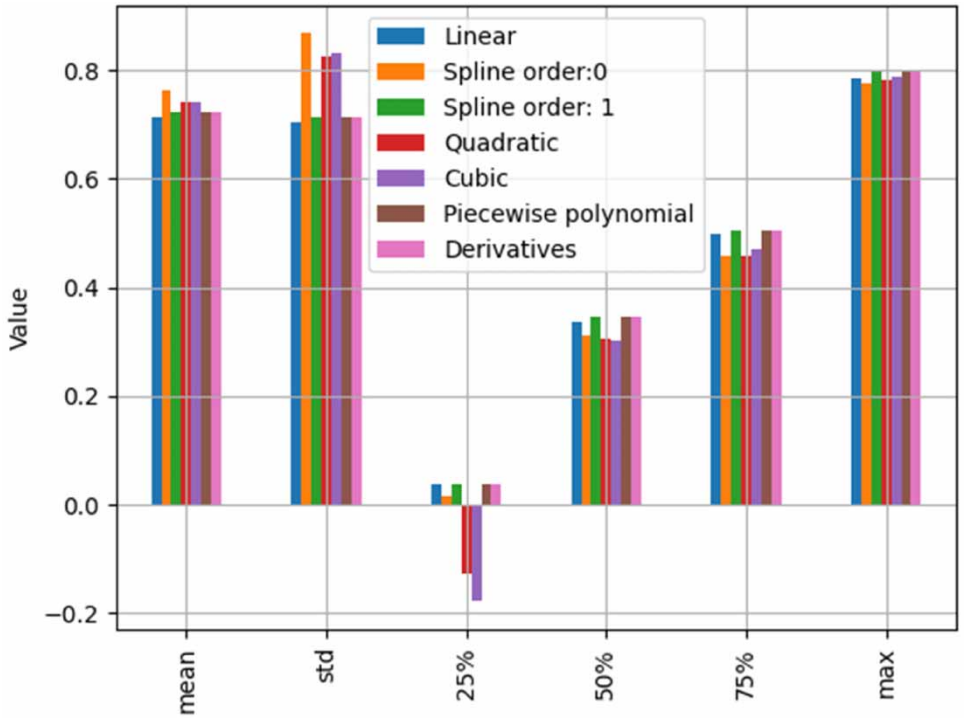


Figure 8 | Basic Model output descriptive statistics.

Figure 9 indicates similar data distributions for the Basic Model outputs when applying different interpolation methods. The median lies in a similar position for all the methods. In addition, the sizes of the box plots are similar, suggesting that the methods may have resulted in a similar output data set when comparing the methods used.

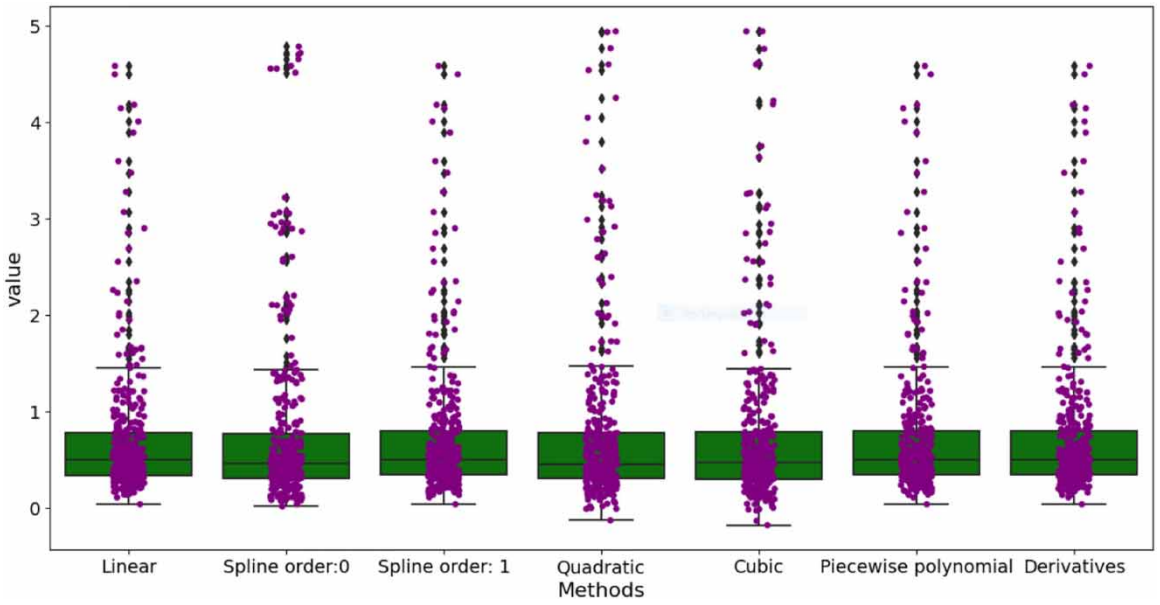


Figure 9 | Basic Model output data.

WASP model

A similar data analysis was performed in application of the WASP model. Figure 10 shows descriptive statistics of the model outputs per interpolation method.

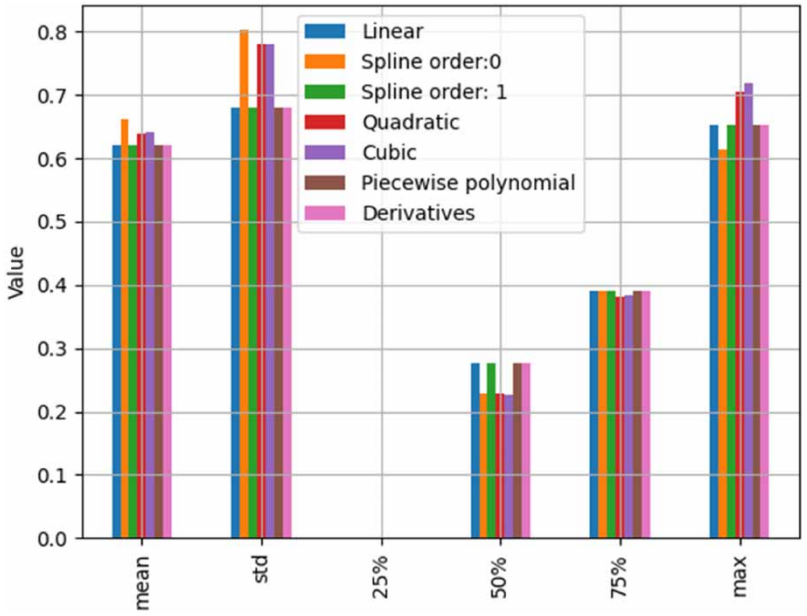


Figure 10 | Descriptive statistics for each interpolation method on the WASP model output.

As was found for the Basic Model results, the WASP model results showed seemingly small variations. The 25% IQR value for this model yielded 0 meaning there was no discernible variability between the methods.

Figure 11 shows model output data distributions to be similar for each of the interpolation methods. This corresponds to the finding made for the Basic Model.

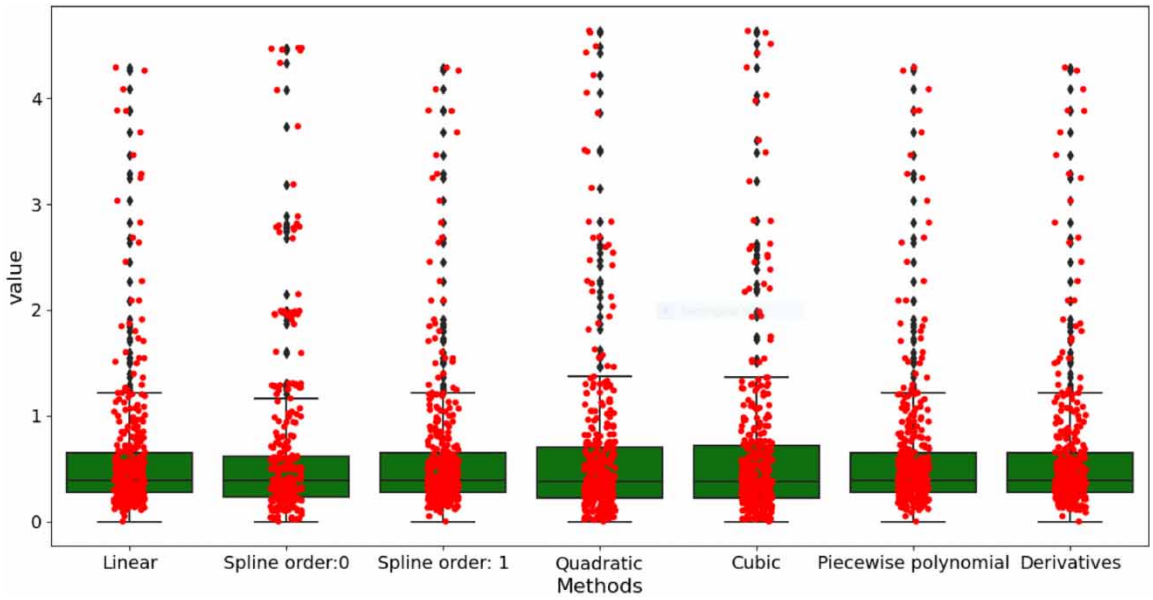


Figure 11 | WASP model box plot showing the data distribution for each method.

Application of artificial gaps

Outputs from the basic and WASP models were compared with downstream measured ammonia concentrations for different sets of input datasets. Figure 12 shows results for the original input data set with no artificial gaps before augmentation. The effect of 10, 25, and 50% are shown in Figures 13–15, respectively.

The model outputs from the Basic Model and WASP produced similar trends for each interpolation method for all input data with varying artificial gaps. The results from the low gaps dataset (10%) and medium gaps dataset (25%) in Figures 9 and 10 infilled the ammonia concentration with a higher value than the actual measured

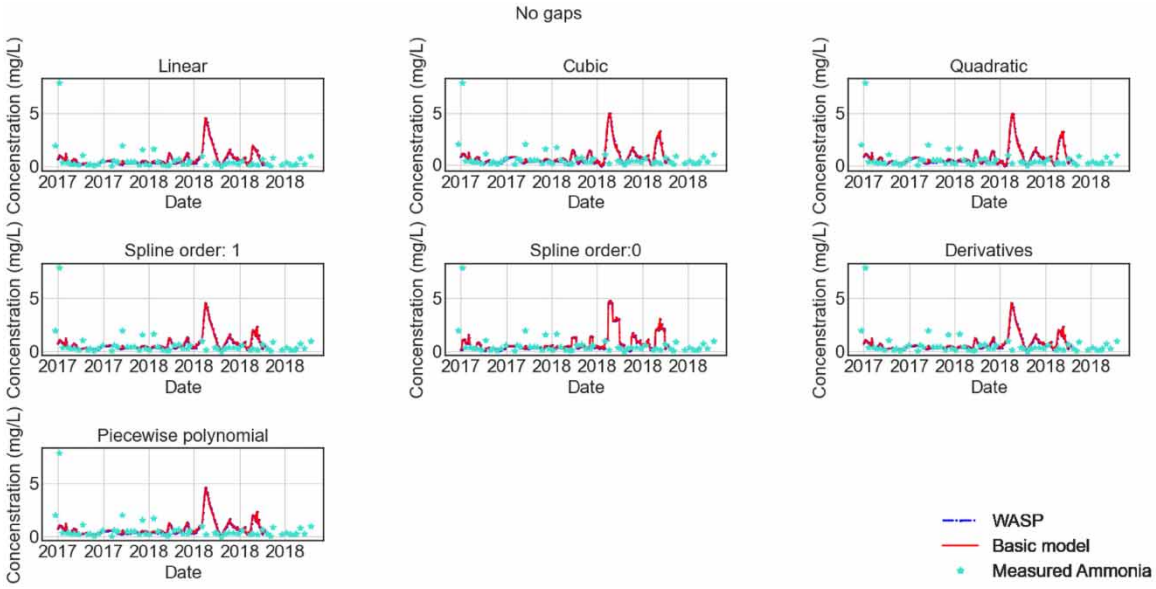


Figure 12 | Model outputs at a downstream location using inputs with no artificial gaps.

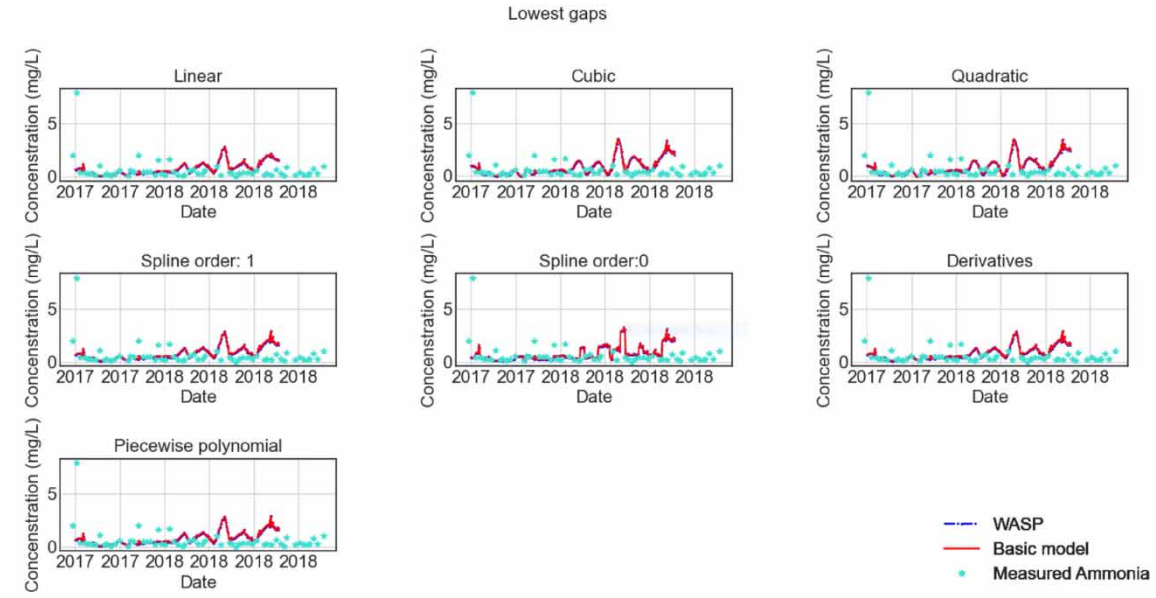


Figure 13 | Model outputs at a downstream location using inputs with low (10%) artificial gaps.

values at the original dataset. This is because the random gaps were introduced at a critical peak towards the end of the time series. The dataset with medium gaps (25%) (Figure 10) resulted in the largest deviation between interpolation method outputs for both models. This indicates that the location of gaps may be of importance as opposed to the level of gap sizes. Furthermore, the results indicate that in the case that random gaps were introduced at critical peaks, the choice of interpolation method may be important. This notion is supported when looking at the case in which high gaps (50%) were introduced, which indicated few differences between the interpolation methods, contrary to intuition.

Further statistical investigation of the simulation results was done to determine and to quantify differences between methods if they exist. The ANOVA test was applied to the Basic Model output data to compare the different outputs when applying the different interpolation methods, with the varying levels of random gaps on the input datasets. Table 4 provides a summary of the test results.

Table 4 lists a small F-value as well as the p-value (PR) > 0.05 for all datasets except for the one with 25% artificial gaps (PR $= 0.0001$), which suggests that for these datasets the differences between the method outputs were

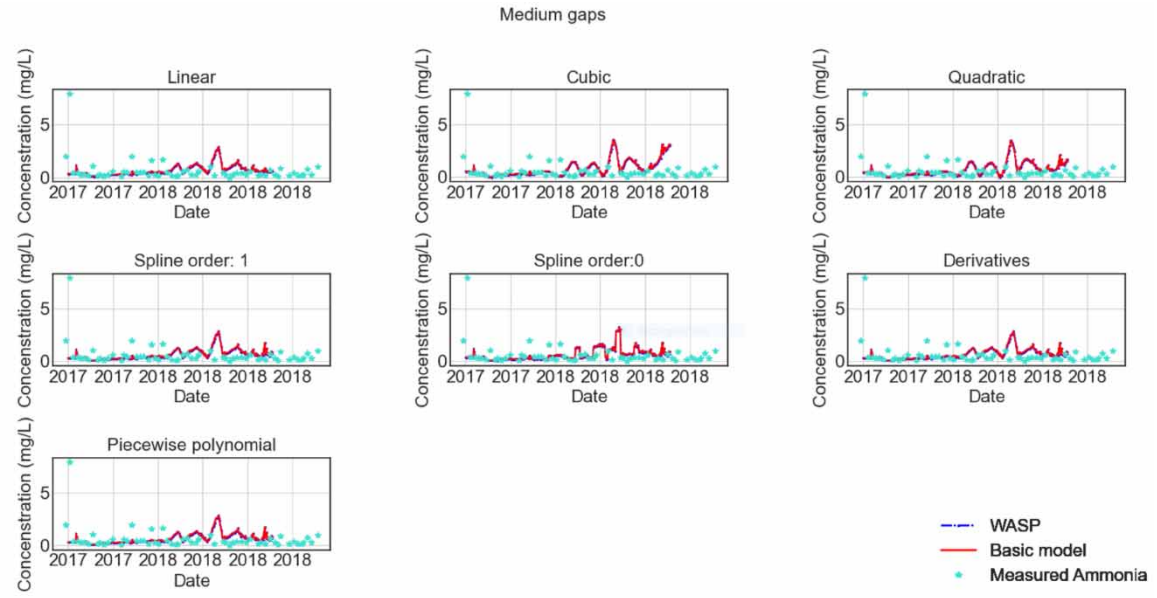


Figure 14 | Model outputs at a downstream location using inputs with medium (25%) artificial gaps.

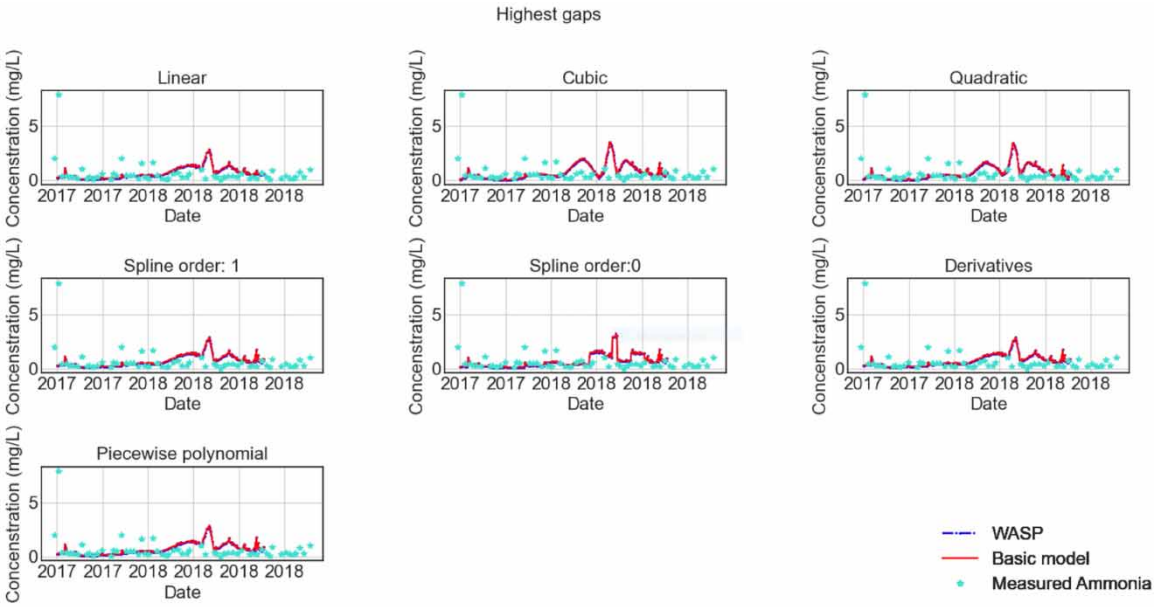


Figure 15 | Model output at a downstream location using inputs with high (50%) artificial gaps.

Table 4 | ANOVA test results

Artificial gaps (%)	ANOVA test	*df*	Sum of squares	Mean square	*F*	PR (>*F*)
0	Methods	6	0.618	0.103	0.1739	0.9839
10	Methods	6	3.9311	0.6552	1.5423	0.1603
25	Methods	6	9.2803	1.5467	4.8975	0.0001
50	Methods	6	1.919	0.3198	0.9581	0.4522

not statistically significant. The contrary result for the case where 25% gaps were included (statistically significant differences in results) may have been due to the introduction of most of the gaps at the peak values in the original dataset. This once more indicated that the location of data gaps is of concern.

Additionally, a *post hoc* (Tukey) test was performed to compare the outputs from the application of the different interpolation methods only on the dataset without artificial gaps. This was to observe if there were any significant differences between the methods themselves. The results of the test are listed in Table 5.

Table 5 | Tukey's honestly significant difference (HSD) test results

Group 1	Group 2	Diff	Lower	Upper	*q*-value	*p*-value
Linear	Spline order:0	0.05	−0.12	0.22	1.25	0.9
Linear	Spline order: 1	0.01	−0.16	0.18	0.29	0.9
Linear	Quadratic	0.03	−0.14	0.2	0.73	0.9
Linear	Cubic	0.03	−0.14	0.2	0.73	0.9
Linear	Piecewise polynomial	0.01	−0.16	0.18	0.29	0.9
Linear	Derivatives	0.01	−0.16	0.18	0.29	0.9
Spline order: 0	Spline order: 1	0.04	−0.13	0.21	0.96	0.9
Spline order: 0	Quadratic	0.02	−0.15	0.19	0.52	0.9
Spline order: 0	Cubic	0.02	−0.15	0.19	0.52	0.9
Spline order: 0	Piecewise polynomial	0.04	−0.13	0.21	0.96	0.9
Spline order: 0	Derivatives	0.04	−0.13	0.21	0.96	0.9
Spline order: 1	Quadratic	0.02	−0.15	0.19	0.43	0.9
Spline order: 1	Cubic	0.02	−0.15	0.19	0.44	0.9
Spline order: 1	Piecewise polynomial	0	−0.17	0.17	0	0.9
Spline order: 1	Derivatives	0	−0.17	0.17	0	0.9
Quadratic	Cubic	0	−0.17	0.17	0.01	0.9
Quadratic	Piecewise polynomial	0.02	−0.15	0.19	0.43	0.9
Quadratic	Derivatives	0.02	−0.15	0.19	0.43	0.9
Cubic	Piecewise polynomial	0.02	−0.15	0.19	0.44	0.9
Cubic	Derivatives	0.02	−0.15	0.19	0.44	0.9
Piecewise polynomial	Derivatives	0	−0.17	0.17	0	0.9

Post hoc tests

The *p*-values for each individual comparison are shown to be above 0.05, which further confirms the rejection of the null hypothesis. These results indicate that there were no statistically significant differences between model outputs when applying the different interpolation methods.

The ANOVA test was similarly applied to the WASP model output data. The results are listed in Table 6.

Table 6 | ANOVA test results for WASP model with artificial gaps

Artificial gaps (%)	ANOVA test	*df*	Sum of squares	Mean square	*F*	PR (>*F*)
0	Methods	6	0.5562	0.0927	0.175	0.9836
10	Methods	6	3.4329	0.5722	1.5476	0.1587
25	Methods	6	8.4919	1.4153	5.0719	0.0000
50	Methods	6	1.7188	0.2865	0.9841	0.4342

The results from this test also support the rejection of the null hypotheses (PR > 0.05) except, once more, for the input dataset with 25% artificial gaps. Therefore, as was found for the Basic Model, differences in the output values when applying the different interpolation methods were not statistically significant for the WASP model. This is further confirmed by the *post hoc* test result listed in Table 7, for the input dataset without gaps.

Method-by-method comparison yielded a rounded-up *p*-value > 0.05 for each case, solidifying the rejection of the null hypothesis. This corresponds to the results from the outputs obtained on the data set generated by the Basic Model.

Table 7 | Tukey's honestly significant difference (HSD) test results

Group 1	Group 2	Diff	Lower	Upper	*q*-value	*p*-value
Linear	Spline order:0	0.04	−0.12	0.2	1.06	0.9
Linear	Spline order: 1	0	−0.16	0.16	0	0.9
Linear	Quadratic	0.02	−0.14	0.18	0.51	0.9
Linear	Cubic	0.02	−0.14	0.18	0.55	0.9
Linear	Piecewise polynomial	0	−0.16	0.16	0	0.9
Linear	Derivatives	0	−0.16	0.16	0	0.9
Spline order: 0	Spline order: 1	0.04	−0.12	0.2	1.06	0.9
Spline order: 0	Quadratic	0.02	−0.14	0.18	0.56	0.9
Spline order: 0	Cubic	0.02	−0.14	0.18	0.51	0.9
Spline order: 0	Piecewise polynomial	0.04	−0.12	0.2	1.06	0.9
Spline order: 0	Derivatives	0.04	−0.12	0.2	1.06	0.9
Spline order: 1	Quadratic	0.02	−0.14	0.18	0.51	0.9
Spline order: 1	Cubic	0.02	−0.14	0.18	0.55	0.9
Spline order: 1	Piecewise polynomial	0	−0.16	0.16	0	0.9
Spline order: 1	Derivatives	0	−0.16	0.16	0	0.9
Quadratic	Cubic	0	−0.16	0.16	0.05	0.9
Quadratic	Piecewise polynomial	0.02	−0.14	0.18	0.51	0.9
Quadratic	Derivatives	0.02	−0.14	0.18	0.51	0.9
Cubic	Piecewise polynomial	0.02	−0.14	0.18	0.55	0.9
Cubic	Derivatives	0.02	−0.14	0.18	0.55	0.9
Piecewise polynomial	Derivatives	0	−0.16	0.16	0	0.9

CONCLUSIONS

The study results indicated that there was no statistically significant difference between the outcomes of each interpolation method applied to the full dataset, however, the introduction of random artificial gaps resulted in significant differences in outcomes between interpolation methods for the case where 25% of the gaps were introduced to the original dataset. Machine learning approaches produced reasonably accurate results. However, upsampling was necessary to obtain the recommended minimum data size, required for learning.

The following was concluded:

- Application of the different interpolation methods applied to input water quality data did not produce statistically significantly different augmented datasets with low gaps.
- Increasing the gaps in the original data sets did not always yield greater differences between augmented datasets for each method.
- The locations of the artificial gaps created statistically significant differences between the augmented datasets for each interpolation method when compared to the effect of high gaps.
- The selected machine learning methods to infill real and artificial gaps were successful in upsampling the original dataset and the dataset with artificial gaps.
- There was no significant difference between the simulated output of WASP and the Basic Model.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Alwosheel, A., Cranenburgh, S. V. & Chorus, C. G. 2018 Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *Journal of Choice Modelling* **28**, 167–182.

Baffaut, C., Dabney, S. M., Smolen, M. D., Youssef, M. A., Bonta, J. V., Chu, M. L., Guzman, J. A., Shedekar, V. S., Jha, M. K. & Arnold, J. G. 2015 Hydrologic and water quality modeling: spatial and temporal considerations. *Transactions of the ASABE* **58**(6), 1661–1680.

Barker, P. M. & McDougall, T. J. 2020 Two interpolation methods using multiply-rotated piecewise cubic hermite interpolating polynomials. *Journal of Atmospheric and Oceanic Technology* **37**(4), 605–619.

Blöschl, G. & Sivapalan, M. 1995 Scale issues in hydrological modelling: a review. *Hydrological Processes* **9**, 251–290.

Chapra, S. 1997 *Surface Water-Quality Modeling*. Waveland Press, Inc., Long Grove, Illinois.

Daggupati, P., Pai, N., Ale, S., Zeckoski, R. W., Jeong, J., Parajuli, P. B., Saraswat, D. & Youssef, M. A. 2015 A recommended calibration and validation strategy for hydrologic and water quality models. *Transactions of the ASABE* **58**(6), 1705–1719.

Dimitriadis, P., Koutsoyiannis, D., Iliopoulou, T. & Papanicolaou, P. 2021 A global-scale investigation of stochastic similarities in marginal distribution and dependence structure of key hydrological-cycle processes. *Hydrology* **8**(59), 1–26.

Harding, W. R. 2015 Living with eutrophication in South Africa: a review of realities and challenges. *Transactions of the Royal Society of South Africa* **70**(2), 155–171.

Kim, J., Seo, D., Jang, M. & Kim, J. 2021 Augmentation of limited input data using an artificial neural network method to improve the accuracy of water quality modeling in a large lake. *Journal of Hydrology* **602**, 126817.

Koutsoyiannis, D., Yao, H. & Georgakakos, A. 2008 Medium-range flow prediction for the Nile: a comparison of stochastic and deterministic methods. *Hydrological Sciences Journal* **53**, 142–164.

László, L. 2005 Cubic spline interpolation with quasiminimal B-spline coefficients. *Acta Mathematica Hungarica* **107**(1–2), 77–87.

Lee, C. J., Hirsch, R. M., Schwarz, G. E., Holtschlag, D. J., Preston, S. D., Crawford, C. G. & Vecchia, A. V. 2016 An evaluation of methods for estimating decadal stream loads. *Journal of Hydrology* **542**, 185–203.

Luna, A. M., Lineros, M. L., Gualda, J. E., Cervera, J. V. G. & Luna, J. M. M. 2020 Assessing the best gap-filling technique for river stage data suitable for low capacity processors and real-time application using iot. *Sensors (Switzerland)* **20**(21), 1–22.

Mehlig, B. 2021 *Machine Learning with Neural Networks, Lecture Notes*. Department of Physics University of Gothenburg.

Moriasi, D. N., Wilson, B. N., Douglas-Mankin, K. R., Arnold, J. G. & Gowda, P. H. 2012 Hydrologic and water quality models: use, calibration, and validation. *Transactions of the ASABE* **55**(4), 1241–1247.

Rozos, E., Dimitriadis, P. & Bellos, V. 2022 Machine learning in assessing the performance of hydrological models. *Hydrology* **9**(5), 1–17.

Sandoval, S., Vezzaro, L. & Bertrand-Krajewski, J.-L. 2016 Gap-filling of dry weather flow rate and water quality measurements in urban catchments by a time series modelling approach. Novatech 2016. In *9th International Conference on Planning and Technologies for Sustainable Urban Water Management*. pp. 1–4.

SAWS 2020 Annual state of the climate of South Africa 2020. *South African Weather Service* **1**(1), 30.

Siauw, T. & Bayen, A. M. 2015 *An Introduction to MATLAB Programming and Numerical Methods for Engineers*. illustrations; 24cm. Academic Press, an imprint of Elsevier, Amsterdam SE – xix. p. 317.

Slaughter, A. R. 2017 Simulating microbial water quality in data-scarce catchments: an update of the WQSAM model to simulate the fate of *Escherichia coli*. *Water Resources Management* **31**(13), 4239–4252.

Slaughter, A. R., Hughes, D. A., Retief, D. C. H. & Mantel, S. K. 2017 A management-oriented water quality model for data scarce catchments. *Environmental Modelling and Software* **97**, 93–111.

The pandas development team 2021 *pandas-dev/pandas*. Pandas. Zenodo. Sebastopol, California.

Tugores, F. & Tugores, L. 2017 Interpolation by derivatives in H∞. *Acta Mathematica Hungarica* **153**(2), 265–275.

Vandebogert, K. 2017 *Method of Quadratic Interpolation*. *PhD Thesis*, University of South Carolina, United States of America.

van Ginkel, C. E. 2011 Eutrophication: present reality and future challenges for South Africa. *Water SA* **37**(5), 693–701.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, I., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F. & van Mulbregt, P. & SciPy 1.0 Contributors 2020 Scipy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261–272.

Vizcaino, I. P., Carrera, E. V., Sanromán-Junquera, M., Muñoz-Romero, S., Rojo-Álvarez, J. L. & Cumbal, L. H. 2016 Spatio-temporal analysis of water quality parameters in Machángara river with nonuniform interpolation methods. *Water (Switzerland)* **8**(11), 1–17.

WWF-SA. 2016 Water: Facts and Futures Rethinking South Africa's Water Future. Report WWF-SA/2016, WWF-SA, Cape Town, South Africa.

Wool, T., Ambrose, R. B., Martin, J. L. & Comer, A. 2020 WASP 8: the next generation in the 50-year evolution of USEPA's water quality model. *Water (Switzerland)* **12**(5), 1–33.

Yang, W., Zhao, Y., Wang, D., Wu, H., Lin, A. & He, L. 2020 Using principal components analysis and IDW interpolation to determine spatial and temporal changes of surface water quality of Xin'anjiang River in Huangshan, China. *International Journal of Environmental Research and Public Health* **17**(2942), 2–14.

Zou, L., Song, L., Wang, X., Weise, T., Chen, Y. & Zhang, C. 2020 A New approach to Newton-type polynomial interpolation with parameters. *Mathematical Problems in Engineering* **2020**, 1–15.

First received 3 March 2022; accepted in revised form 9 November 2022. Available online 18 November 2022